Phishing in Style: Characterizing Phishing Websites in the Wild

David Hasselquist*†, Elsa Kihlberg Gawell*, Axel Karlström*, Niklas Carlsson*
*Linköping University, Sweden
†Sectra Communications, Sweden

***Abstract*—The prevalence of phishing domains is steadily rising as attackers exploit toolkits to create phishing websites. As web development expertise is no longer a prerequisite, phishing attacks have become more widespread, outpacing many existing detection methods. Developing novel techniques to identify malicious domains is crucial to safeguard potential victims online. While most current methods emphasize the visual aspects of phishing websites, in this paper, we investigate the underlying structure by collecting data on style sheets and certificates from both verified phishing domains and benign domains. Using a token-based similarity algorithm, we group the phishing domains into three categories and identify shared characteristics of these domains. Our work demonstrates the feasibility of using structural similarities to identify a website created using a phishing kit. By employing such detection, users would be able to browse the web with a reduced risk of falling victim to malicious activities.**

I. Introduction

Phishing attacks have become an increasingly prevalent threat on the internet, with attackers manipulating victims into providing sensitive information through malicious domains that visually imitate trusted entities [1]. The increasing use of phishing kits, allowing attackers with little to no web design experience to create phishing domains, has contributed to the rise of these attacks. These kits simplify the process of crafting visually deceptive domains, leading to an increasing number of phishing websites that are difficult for users to identify.

However, as phishing kits are usually designed to create a variety of web pages similar to several target websites, the structural similarities of created web pages are often more distinct than visual ones. The varying appearances of the created websites make them difficult to detect with visual-based detection methods. This makes the analysis of the phishing domain infrastructure particularly relevant, as it reveals information that cannot be detected through visual methods alone. Even though structural-based detection methods have proven to be effective against phishing kits [2], phishing kits are fast adapting to such detection methods.

In this paper, we take a closer look at the underlying structure of domains by studying Cascading Style Sheets (CSS) and X.509 certificates issued by a Certificate Authority (CA). First, we describe our dataset and the differences in our data collection process for phishing and benign domains of varying popularity (§II). Second, we study the different characteristics of these domains, focusing on loading time, certificates, and style sheets (§III). Third, we describe our token-based similarity algorithm that groups phishing domains based on their style sheets and provide insights into their structural similarities and differences (§IV). Finally, we briefly describe related works (§V) before concluding the paper (§VI).

978-3-903176-58-4

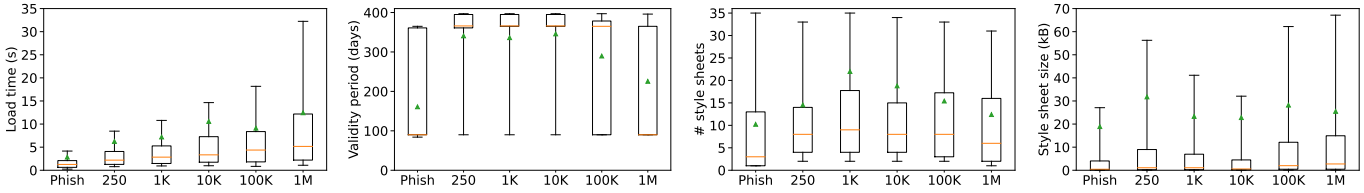
II. Dataset

Due to the differences in phishing and benign domains, we use a separate data collection process for each category.

Phishing domains: For our phishing domain dataset, we use domains obtained from PhishTank [3] where users can submit suspected phishing domains along with their suspected target domains. Verified and active phishing domains are updated every hour. Additionally, other users can verify these domains, contributing to phishing prevention efforts. We select this particular database due to its accessibility, constant updates, and extensive list of domains [4]. Considering their typically short lifespan, we continually acquire the phishing domains and immediately process them. This ensures the maximum number of active domains during our data collection and provides us with up-to-date samples of active phishing websites, rather than relying on outdated or offline domains in older databases. Our phishing domain dataset consists of 4,422 domains.

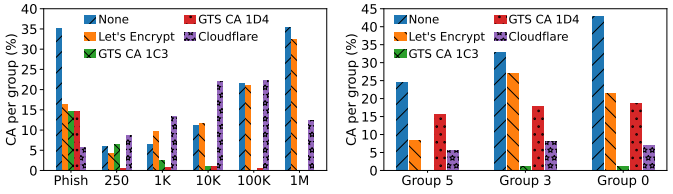
Benign domains: For our benign domain dataset, we use a subset from the top one million most popular domains on the Tranco list [5]. This list is derived based on the 30-day averages of four major domain lists: Alexa, Majestic, Cisco Umbrella, and Quantcast [5]. To limit our dataset and enable comparison with different domain popularities, we select the first 250 domains (ranks 1–250) and the last 250 domains from each magnitude sample (i.e., ranks 751–1,000, 9,751–10,000, 997,511–100,000, and 999,751–1,000,000). We denote these subsets as Tranco 250, 1K, 10K, 100K, and 1M, respectively. In total, our benign domain dataset consists of 1,250 domains.

Data collection: For both our phishing and benign dataset, we use Selenium [6] to crawl each domain and collect the domain and website information. At a high-level, for each domain, we collect (1) timestamps corresponding to different events (e.g., loading times), (2) certificate information (e.g., CA and validation time), (3) source URL for any style sheets, (4) length of each style sheet, and (5) class names in each identified style sheet. We collect data from both active and inactive stylesheets. However, due to their extensive length, we restrict our collection of stylesheet information to only include class names contained within them. Additionally, we exclude non-responsive domains from our dataset.



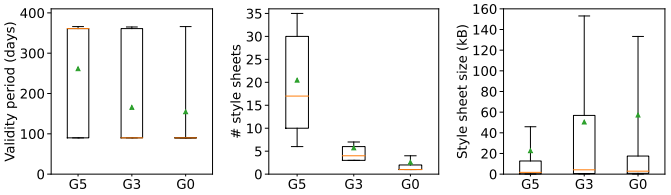
(a) Webpage load time (b) Certificate validity period (c) Number of CSS (d) CSS size

Fig. 1. Load time, certificate validity period, CSS amount, and CSS size per domain group (phishing domains and benign domains of different rank popularity).



(a) Per domain category (b) Per CCS group

Fig. 2. Certificates issued per CA (top-5 CA cases in phishing dataset).



(a) Cert. validity period (b) Number of CSS (c) CSS size

Fig. 3. Certificate validity period, CSS amount, and CSS size per CSS group.

III. Phishing vs. Benign Domains

Using our dataset of phishing domains and benign domains, we next present a high-level comparison.

Loading time: Figure 1(a) shows a box plot of the load time for different domain groups (i.e., phishing domains and benign domains of different popularity as ranked on the Tranco list). Here, we show the 10^{th} percentile (bottom marker), 25^{th} percentile (bottom of box), median (middle marker), 75^{th} percentile (top of box), 90^{th} percentile (top marker), and average (triangle). We observe the loading times for phishing domains to be shorter than those for benign domains, and that the loading time increases as the domain popularity drops. The faster loading time of phishing domains can be explained by them usually serving only one purpose with a small content size, without the need for extra features or 3rd-party content.

Certificate authority: Figure 2(a) shows the usage of the five CAs most commonly used by the phishing domains; both for the phishing domains and the benign domains with different popularity rankings. For the phishing domains and the Tranco-1M group, 35% are missing a certificate (None). The most common certificate issuer for the phishing domains is Let's Encrypt, and we see its increasing popularity among the benign domains as the rank drops. For the phishing domains, we also observe many containing Google-issued certificates (GTS). As we will see later, this is due to many of them redirecting to Google.com.

Certificate validity: Figure 1(b) shows a box-plot of the certificate validity period for each domain group. Again, we see similarities between the phishing domains and the Tranco 1M group, with a median validity period of 90 days (enforced by Let's Encrypt). The use of Let's Encrypt (and also 90 days validity periods) for phishing domains is not surprising. Since phishing sites tend to be short-lived, certificates obtained through Let's Encrypt may be preferred as they are free and can be obtained using an automated process. For other groups, validity periods of approximately one year dominate, although we also observe the 10^{th} percentile to include 90 days.

Number of style sheets: Figure 1(c) shows a box plot of the number of style sheets observed for each domain group. Among all groups, we observe the phishing domains to use the least number of style sheets. This is expected due to their simplicity, with a sole purpose of collecting sensitive data, also reflected in the low loading time (Figure 1(a)).

Style sheets size: Figure 1(d) shows the distribution of the style sheet sizes for each domain group. Here, we see that the phishing domains are more similar to the intermediate popular domains (1K–10K), possibly indicating that phishing domains more often target the more popular domains, and that the style sheet sizes increase slightly for the less popular domains (100K–1M). For all domain groups, we observe a low median (inside the boxes) and a large average (far outside the boxes), indicating the presence of outliers with significantly larger sizes, often due to bundled style sheets.

IV. CSS Similarity

Token-based similarity algorithm: To find structural similarities in style sheets, we focus on the class names from each style sheet present on the domains in our phishing dataset and use a token-based similarity algorithm. For each domain, we check if any style sheet is similar to a style sheet of another domain by comparing their Jaccard distances. If the distance is greater than 0.9 (i.e., if class names are at least 90% similar), we consider these to be similar. If at least five style sheet matches between two domains, we place the matching domains exclusively into a group denoted as G5. For the remaining domains (i.e., those with fewer than five matches), we run the same algorithm using a threshold of three matches and place these matching domains exclusively into G3. The remaining domains with fewer than three matches are denoted as G0. This gives us three groups of phishing domains containing different number of style sheet matches within the dataset.

Out of the domains our phishing dataset, G5 contains 1,305 domains (29.5%), G3 731 domains (16.5%), and G0 2,386 domains (54.0%). However, upon closer inspection in G5 and G3, we exclude 609 and 176 domains, respectively, as they redirect to Google.com or contain an error page. We next present a structural analysis of the three CSS groups.

Certificate authority: Figure 2(b) shows the five most common CAs for the different CSS groups. After removing the domains that redirect to Google.com, we now see a low certificate usage of GTS CA 1C3. Comparing G5 and G0, we observe that G0 has a lower usage of certificates but a significantly higher usage of Let's Encrypt.

Certificate validity: Figure 3(a) shows the certificate validity period for the CSS groups. While we observe that all three groups use a similar range of certificate validity times (90–365 days), we see a trend of larger validity times with increasing matches of domains (from G0 to G5).

Number of style sheets: Figure 3(b) shows the distribution of the number of style sheets for the CSS groups. Here, we see large variances, showing that the number of style sheets on phishing domains are not evenly distributed. Still, the tight distribution in G0 and G3 is interesting.

Style sheet size: Figure 3(c) shows the distribution of the style sheet sizes for the CSS groups. Here, most notably is G3, having a large variance and the 10^{th} percentile reaching over 150kB. We observe the shortest style sheets to be used in G5. Again, for the CSS sizes, a low median and a large average is the result of significant outliers.

In summary, when comparing G5 to G3, we observe two distinct approaches in their use of style sheets. G5 employs a larger number of shorter style sheets, while G3 opts for fewer, but more extensive ones. The style sheets of G5 suggest that importing style sheets is a prevalent practice among the domains within this group. Given that phishing domains typically lack substantial content or functionality, generating such a high number of style sheets seems counterintuitive. On the other hand, the style sheets of G3 reflect a more comprehensive approach, where only a select few style sheets are imported, and the rest of the styling is consolidated into one or two primary style sheets. This pattern is more indicative of phishing kits being employed, where the large size of the style sheets allows for the adaptation of structural appearances to suit various target domains.

Subset analysis: We next look closer at the largest subset (224 domains) of G5, containing matches with each other. Here, all domains are subdomains of "weebly.com," a free web hosting service featuring a drag-and-drop builder and a certificate issued by DigiCert. The subset contains 49 unique style sheets, with varying lengths from 44 to 215,000 characters (majority fewer than 7,000 characters). While our method in practice may have a high false positive rate and benign domains may also be identified, this demonstrates the ability to use CSS data to differentiate and categorize domains created using the same website-building tool (e.g., phishing kits).

For the largest subset (66 domains) of G3, most domains are primary domains. With 14 unique style sheets, this subset stands out with style sheet lengths between 20,000 and 150,000 characters. Here, most style sheets are not minimized, and websites only use a small part of them, with the ability to display visually different pages using the same style sheet. The lack of style sheets minimization benefits phishing kits, as they easily can create multiple visually diverse websites.

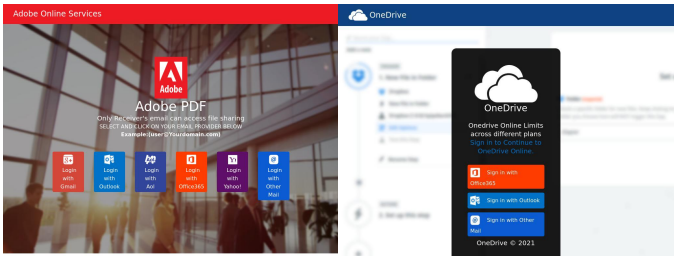


Fig. 4. Screenshot of two visually different phishing domains using similar style sheets, detected and grouped using our token-based similarity algorithm.

To demonstrate the effectiveness of identifying two visually different websites that use the same large style sheet, Figure 4 shows two user-submitted screenshots from PhishTank for two domains in this subset. In this case, our method effectively groups the two phishing domains based on their style sheets, despite their apparent visual differences. This suggests that employing similar approaches to identify and compare structural similarities could prove useful in detecting phishing domains where visual-based detection might fail.

V. Related Work

While much previous works focus on identifying and categorizing phishing domains using visual appearances like analyzing URL [7]–[12], visual similarities [13]–[17] or DNS information [16], others focus on the website structure [2], [12], [18] or phishing kits [19]–[23]. Although challenging, some works also use certificate information and try to separate phishing domains from benign domains [21], [24]. Unlike visual approaches, and structural approaches focusing on individual objects, we have taken an alternative approach to study structural similarities based on the style sheet data and certificate data. Similar to our approach, many works [1], [2], [4], [8], [11]–[14], [16], [19]–[21], [24] use Phishtank [3] due to its availability and extensive list of updated domains.

VI. Conclusion

In this paper, we have studied the feasibility of using structural similarities to detect phishing websites. By analyzing the characteristics of the style sheets and certificates of both phishing and benign domains, we show that phishing domains have distinctive characteristics, such as shorter loading times, fewer style sheets, and shorter validity times for certificates. Using a token-based similarity algorithm to group phishing domains based on their style sheets, we highlight different traits among these groups, showing the possible use of phishing kits in one of the groups. Our results suggest that such structural-based detection methods can play an important role in detecting and preventing phishing attacks, complementing existing visual-based detection methods.

ACKNOWLEDGMENT

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

[1] P. Zhang, A. Oest, H. Cho, Z. Sun, R. Johnson, B. Wardman, S. Sarker, A. Kapravelos, T. Bao, R. Wang *et al.*, "Crawlphish: Large-scale analysis of client-side cloaking techniques in phishing," in *IEEE Symposium on Security and Privacy (S&P)*, 2021.

[2] S. Tanaka, T. Matsunaka, A. Yamada, and A. Kubota, "Phishing site detection using similarity of website structure," in *Proc. IEEE Dependable and Secure Computing (DSC)*, 2021.

[3] Cisco Talos Intelligence Group, "Phishtank," https://phishtank.com/, 2023.

[4] S. Bell and P. Komisarczuk, "An analysis of phishing blacklists: Google Safe Browsing, OpenPhish, and PhishTank," in *Proc. Australasian Computer Science Week (ACSW)*, 2020.

[5] V. L. Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," 2018.

[6] "Selenium," https://www.selenium.dev/, 2023.

[7] A. Blum, B. Wardman, T. Solorio, and G. Warner, "Lexical feature based phishing url detection using online learning," in *Proc. ACM CCS Workshop on Artificial Intelligence and Security (AISec)*, 2010.

[8] A. Anand, K. Gorde, J. R. A. Moniz, N. Park, T. Chakraborty, and B.-T. Chu, "Phishing url detection with oversampling based on text generative adversarial networks," in *Proc. IEEE Big Data*, 2018.

[9] K. Althobaiti, G. Rummani, and K. Vaniea, "A review of human-and computer-facing url phishing features," in *Proceedings IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2019.

[10] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in *Proc. ACM CCS Workshop on Recurring malcode (WORM)*, 2007.

[11] S. Marchal, K. Saari, N. Singh, and N. Asokan, "Know your phish: Novel techniques for detecting phishing sites and their targets," in *Proc. IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2016.

[12] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE access*, 2019.

[13] J. Mao, P. Li, K. Li, T. Wei, and Z. Liang, "BaitAlarm: Detecting phishing sites using similarity in fundamental visual features," in *Proc. IEEE Intelligent Networking and Collaborative Systems (INCoS)*, 2013.

[14] S. Abdelnabi, K. Krombholz, and M. Fritz, "Visualphishnet: Zero-day phishing website detection by visual similarity," in *Proc. ACM Computer and Communications Security (CCS)*, 2020.

[15] J. Mao, W. Tian, P. Li, T. Wei, and Z. Liang, "Phishing-alarm: robust and efficient phishing detection via page component similarity," *IEEE Access*, 2017.

[16] K. Tian, S. T. Jan, H. Hu, D. Yao, and G. Wang, "Needle in a haystack: Tracking down elite phishing domains in the wild," in *Proc. Internet Measurement Conference (IMC)*, 2018.

[17] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong, "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages," in *Proc. USENIX Security*, 2021.

[18] Y. Pan and X. Ding, "Anomaly based web phishing page detection," in *Proc. Annual Computer Security Applications Conference (ACSAC)*, 2006.

[19] X. Han, N. Kheir, and D. Balzarotti, "Phisheye: Live monitoring of sandboxed phishing kits," in *Proc. ACM Computer and Communications Security (CCS)*, 2016.

[20] A. Oest, Y. Safaei, A. Doupé, G.-J. Ahn, B. Wardman, and K. Tyers, "Phishfarm: A scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists," in *IEEE Symposium on Security and Privacy (S&P)*, 2019.

[21] B. Kondracki, B. A. Azad, O. Starov, and N. Nikiforakis, "Catching transparent phish: Analyzing and detecting mitm phishing toolkits," in *Proc. ACM Computer and Communications Security (CCS)*, 2021.

[22] A. Oest, Y. Safaei, P. Zhang, B. Wardman, K. Tyers, Y. Shoshitaishvili, A. Doupé, and G.-J. Ahn, "Phishtime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists," in *Proc. USENIX Security*, 2020.

[23] H. L. Bijmans, T. M. Booij, A. Schwedersky, A. Nedgabat, and R. van Wegberg, "Catching phishers by their bait: Investigating the dutch phishing landscape through phishing kit detection." in *Proc. USENIX Security*, 2021.

[24] U. Meyer and V. Drury, "Certified phishing: taking a look at public key certificates of phishing websites," in *Proc. Symposium on Usable Privacy and Security (SOUPS)*, 2019.